TRANSFER LEARNING ON ULTRASOUND SPECTROGRAMS OF RF SIGNAL FOR DIAGNOSING NONALCOHOLIC FATTY LIVER DISEASE

By

Shi Chen

Senior Thesis in Electrical Engineering

University of Illinois at Urbana-Champaign

Advisors: William D. O'Brien, Jr. and Aiguo Han

May 2019

Abstract

Purpose: Nonalcoholic fatty liver disease (NAFLD) is a very common liver disease and affects 20-30% of the world population. The gold standard for diagnosing NAFLD is the liver biopsy. The purpose of this thesis is to develop and evaluate a 2D convolutional neural network (CNN) which uses the ultrasound spectrograms of radiofrequency (RF) signals for diagnosing NAFLD. Spectrogram of an RF signal shows the frequency spectrum of the RF signal as it varies with time.

Methods: 204 research participants with and without NAFLD took both ultrasound imaging and MRI. The results given by the MRI-proton density fat fraction (PDFF) were used as the reference (NALFD defined as PDFF≥5%). For ultrasound images from all patients, a fixed region-of-interest (ROI) was chosen to maximize the portion of the liver in the ROI. The time-gain-compensation (TGC) was removed from the RF signals and RF signals were downsampled and normalized. For each RF signal, a spectrogram with size 32×32 was generated and the log-transformed absolute values of spectrograms were used as inputs for the 2-D CNN model. A transfer learning approach was used, where a pretrained VGG-19 model was used as the base model and was fine-tuned by training the last few convolution layers and fully-connected layers to diagnose NAFLD. The hyperparameter optimization was done on the training set by a grid search to optimize the number of trainable convolution layers and the number of nodes in the first fully-connected layer.

Results: The model achieved a sensitivity of 94.3%, a specificity of 90.6% and an accuracy of 93.1%s for diagnosing NAFLD on the test set.

Conclusions: It is feasible to develop a 2D model from transfer learning on ultrasound spectrograms of RF signal to accurately diagnose NAFLD. The proposed approach provides a noninvasive method to evaluate the fatty liver disease.

Subject Keywords: ultrasound imaging; radiofrequency signal; spectrograms; deep learning; transfer learning; NAFLD

Contents

1. Introduction

Nonalcoholic fatty liver disease (NAFLD) is a very common liver disease which is believed to affect one-third of adults in the US [1]. There are two types of NAFLD: the first type is nonalcoholic fatty liver (NAFL), and the second type is nonalcoholic steatohepatitis (NASH). NAFL is a condition in which only fat is built up in the liver, whereas NASH is a condition in which liver cells are damaged in addition to the built-up fat [1]. Without proper treatment, NASH could progress to liver fibrosis, cirrhosis, and even liver cancer [2]. Diagnosing and treating NAFLD early may stop the progression of the disease and reverse the change in the liver which NAFLD causes.

The current gold standard for clinical diagnosis of NAFLD is liver biopsy, which is an invasive procedure with some potential severe complications [3]. Noninvasive methods to diagnose NAFLD include MRI proton density fat fraction (MRI-PDFF), conventional B-mode ultrasound imaging and quantitative ultrasound (QUS). MRI-PDFF has the advantage of highly accurate result for fat quantification [4-5] but at a very high cost. Conventional B-mode imaging of the liver is inexpensive and widely available, but the inter-sonographer and inter-transducer agreements are relatively low [6]. QUS can accurately predict NAFLD by using parameters such as attenuation coefficient (AC) and backscatter coefficient (BSC) [7]. However, physical phantoms are required to get AC and BSC. Diagnosing NAFLD noninvasively remains a challenging problem.

In ultrasound imaging, the knowledge of tissue properties comes from the analysis of ultrasound radiofrequency (RF) signal. B-mode ultrasound images are constructed from the envelope function of the RF signal and QUS parameters AC and BSC are calculated from the power spectrum of the RF signal. Quantitative analysis of RF signals has yielded substantial results in characterizing coronary plaques [8], infarcted myocardium [9], breast tumor [10], and prostate cancer [11]. These studies all used the RF signal in time-domain. However, frequency-dependent information in the RF signals is shown to be highly associated with the sub-wavelength microstructure of the tissue [12].

To utilize the RF signal in the frequency domain, this study proposed to use transfer learning to provide a phantom-free ultrasound approach to diagnose NAFLD. Transfer learning is a critically important and useful technique of deep learning. Transfer learning is a machine learning technique that transfers the knowledge learned in one domain for some general task from a large amount of labeled data to another specific domain for a different task where the number of labeled data is much smaller [13]. Transfer learning does not need many labeled data to develop a convolutional neural network (CNN) from scratch and can reuse the pretrained CNNs with some modifications. Transfer learning has also been shown to improve convergence speed in medical imaging applications [14]. This study fine-tuned and evaluated a two-dimensional (2D) pretrained CNN model for NAFLD diagnosis using the MRI-PDFF as the reference.

2. Literature Review

Several studies [15-16] have used transfer learning on ultrasound B-mode images to diagnose NAFLD. Byra et al. [15] used the pretrained Inception-ResNet-v2 model for high-level feature extraction in liver B-mode ultrasound images and the support vector machine to classify images. The reference used in [15] is the liver biopsy, and fatty liver is defined as that with fat fraction larger than 5%. The dataset used in [15] contained 55 severely obese patients and 38 of them had fatty liver. The images were acquired by GE Vivid E9 Ultrasound System and 10 images were acquired for each patient. Each whole B-mode image was used as input and it did not require choosing a region of interest. The model achieved a sensitivity of 100%, a specificity of 88.2% and an accuracy of 96.3%. Reddy et al. [16] used VGG-16 model as the base model and fine-tuned the last three convolution layers and two fully-connected layers. Their dataset contained 81 normal liver images and 76 fatty liver images and the images were labeled by the radiographers. They achieved a sensitivity of 95%, a specificity of 85% and an accuracy of 90%.

The results of these studies demonstrate the fact that transfer learning on ultrasound B-mode images is able to provide useful information for diagnosis purposes. These studies essentially utilized the envelope function of RF signal in time-domain. However, RF signals have both time-dependent information and frequency-dependent information. As frequency-dependent information is associated with the microstructure of the tissue, quantitative analysis of frequency-domain RF signal could yield useful results. Therefore, this study focused on analyzing the frequency-domain RF signal by using transfer learning.

3. Description of Research Results

3.1 Dataset description

The dataset used in this study contains data from 204 participants. These 204 participants took ultrasound imaging and MR imaging on the same day. For each patient, 10 frames were acquired during the ultrasound imaging. MR imaging gave the fat fraction of each segment among nine segments of the liver and the average fat fraction over nine segments. The average fat fraction given by the MR scanning was used as the "ground truth" of the liver fat fraction of each patient in this study.

Among 204 patients, 64 were healthy and the remaining 140 had NAFLD. The 204 patients were equally divided into training set and test set. Each set was composed of 32 healthy patients and 70 patients with NAFLD. The RF data used in this study were acquired from Siemens ultrasound system by using transducer 4C1 (the nominal center frequency and bandwidth are 3MHz and 1 – 4 MHz, respectively).

3.2 Data preprocessing

Before using the RF data in the deep learning model, several processing steps were performed on the RF data of all patients. The first step was to downsample the RF data by a factor of 4 so that the number of samples of a gated RF signal was reduced from 4096 to 1024. The primary reason for downsampling was to make the training and testing processing faster. The original sampling frequency was 40 MHz and the new sampling frequency was 10 MHz after downsampling. The downsampling factor 4 was chosen because the data after downsampling still satisfied the Nyquist criterion. The second step was to remove the time gain compensation (TGC) from the RF data to get raw RF data because the ultrasound machine applied the TGC to account for the signal attenuation. The third step was to normalize the RF data such that the mean of RF data is 0 and the standard deviation of RF data is 1. This step was performed only for the purpose of weight initialization of the deep learning model. The fourth step was to fix the ROI such that the majority of the tissue within the ROI is liver. The reason for choosing such an ROI was to minimize the RF signal from other tissue.

The spectrogram of the RF signal was used as the input for the deep learning model. The reason for using spectrogram of RF signal rather that B-mode images was that the spectrogram of RF signal contained a significant amount of information which the B-mode imaged did not have because the envelope detection discards a lot of information. And the frequency information of RF data is highly associated with the acoustic properties of tissue and may be useful for tissue characterization. The spectrogram of RF data was essentially the short-time Fourier transformation of the RF data and therefore retained the important frequency information in the RF data. The spectrograms were generated by using a built-in function called spectrogram in MATLAB and the size of spectrogram is 32×32. The spectrogram of one ultrasound RF signal from a fatty liver is shown in Figure 1.

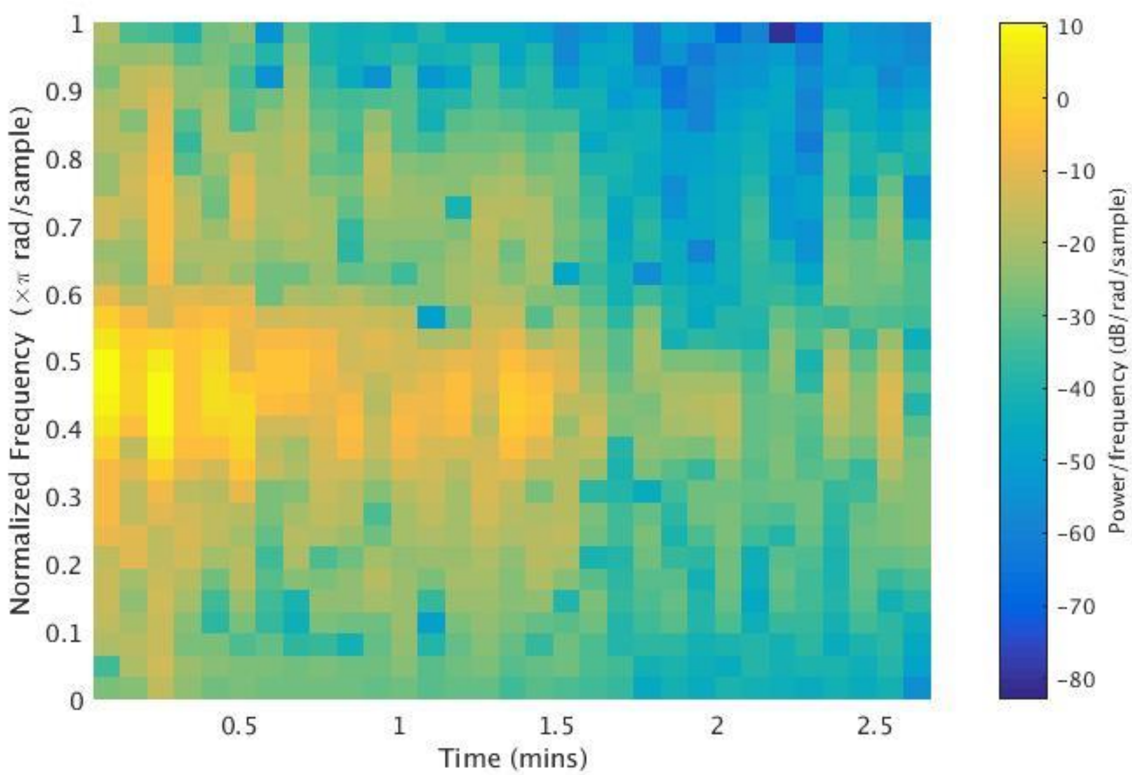

Figure 1 The spectrogram of one ultrasound RF signal from a fatty liver. The window used is rectangular window with length 32, the samples of overlap is 0 and the sampling points for compute discrete Fourier Transform is 62.

The absolute values of the spectrograms were used as the inputs for the model. Since the range of absolute value of spectrogram was different for each A-line and each participant, it is important that all images should have a consistent dynamic range. Therefore, the 20log10 of the maximal absolute value of spectrograms from all patients was set as high limit and a dynamic range of 60dB was chosen. The resulting image was uniformly quantized with 256 levels and the parula colormap was used to assign the color to three channels according to the intensity of each pixel. Parula colormap is the default colormap which defines the colors for graphic objects in MATLAB.

3.3 2-D Deep learning with transfer learning

In this study, deep learning models were developed and used to perform a binary classification between participants with no NAFLD and participants with NAFLD. The model was optimized by using 2-fold cross-validation on the training set to select the best hyperparameters and the optimal model was tested on the test set.

Transfer learning was used to develop the 2D model because the number of available labeled data was small and it is hard to start from scratch to develop a model with many layers. For choosing the model for transfer learning with ultrasound images, there were few hints from the literature. Tajbakhsh et al. [17] used the AlexNet with CT images in pulmonary embolism detection and [18] used VGG with X-ray, MRI, and CT images for medical image classification. In this study, the VGG-19 network was used in transfer learning. VGG-19 model consists of 16 convolution layers and three fully-connected layers and has more than 10 million parameters. VGG-19 has been trained on more than a million images from ImageNet database and can classify images into 1000 different categories.

To use VGG-19 as the model of transfer learning, the first 16 convolution layers and their weights are preserved as the pre-trained model because for the convolution neural network, the initial layers are used to extract simple features while the top fully-connected layers are used to extract abstract and task-specific features [13]. The three fully-connected layers were replaced by two fully-connected layers to reduce mode capacity. Activation functions are functions which computes the weighted sum of input and biases in a model and commonly used activation functions are softmax function, hyperbolic tangent function and rectified linear unit (ReLU) function [19]. The second fully-connected layer used the softmax as the activation function and had 2 nodes because this layer estimates the pseudo probabilities of two classes. The first fully-connected layer used the ReLU as the activation function because ReLU is a faster learning activation function with better performance and generalization [19]. The number of nodes in the first fully-connected layer was one hyperparameter to be optimized and the number of trainable convolution layers was the other hyperparameter to be optimized. Both hyperparameters were determined during the hyperparameter optimization.

Hyperparameter optimization is an optimization procedure which finds the optimal hyperparameters of model which optimizes the performance of the model. The hyperparameter optimization was achieved by using the grid search. The average patient-level validation accuracy of 2-fold cross validation was calculated as the metrics. For each hyperparameter pair and each fold, the model was trained for 20 epochs. The pair which gave the highest average validation accuracy of last 5 epochs was used as the optimal hyperparameters for the final model. Then the final model was tested on the test set and the patient-level test accuracy, specificity and sensitivity were calculated to evaluate the performance of the model.

To get a better understanding of the mechanism behind the deep learning approach, the model was trained and tested on two additional types of inputs with different hyperparameter pairs. The first type of input was constructed by using only the phase information of the spectrogram of the ultrasound RF signal. The phases were uniformly quantized with 256 levels and a colormap called parula was used to assign the color to three channels according to the intensity of each pixel. The second type of input was constructed by combing the absolute value of spectrogram, log-compressed absolute value of spectrogram and the phase information of the spectrogram. The absolute value was put in the R-channel, the log-compressed absolute value was put in the G-channel and the phase information was put in the B-channel. For each type of input, the final model was trained on the whole training set and tested on the test set and the patient-level accuracy was calculated.

3.4 Results

3.4.1 Hyperparameter optimization

For each hyperparameter pair, the average of 2 folds patient-level validation accuracies were calculated. The average validation accuracy of the grid search are in Table 1.

Table 1 Average Validation Accuracy of Grid Search

		Number of nodes in the first fully-connected layer				
		32	64	128	0256	512
Number of trainable convolution layers	0	90.2%	90.0%	89.8%	89.8%	90.4%
	1	91.8%	92.4%	92.8%	92.6%	92.2%
	2	90.6%	91.2%	91.2%	91.4%	91.2%

When the number of trainable convolution layers is fixed at any nonzero value, the average validation accuracy initially increases and then decrease as the number of nodes in the first fully-connected layer increases. When the number of nodes in the first fully-connected layer is fixed, the average validation accuracy increases and then decreases as the number of trainable convolution layers increases.

When the number of nodes in the first fully-connected layer is 128 and the number of trainable convolution layers is 1, the average validation accuracy is highest in table 1. Therefore, the optimal number of nodes in the first fully-connected layer is 128 and the optimal number of trainable convolution layers is 1 for this model on this training set.

3.4.2 Testing result

To assess the performance of this model, the optimal model was trained on the entire training set for 20 epochs and then tested on the test set. The patient-level predictions given by the VGG-19 model were shown in Table 2.

Table 2 Predictions from VGG-19 Model

		Truth condition	
		NAFLD	**Controls**
Test Result	**Positive**	29	4
	Negative	3	66

Among 102 patients in the test set, the model correctly classified 95 of them and the accuracy was 93.1%. For 32 patients with NAFLD, the model correctly classified 29 of them and gave the specificity of 90.6%. For 70 patients without NAFLD, the model correctly classified 66 of them and gave the sensitivity of 94.3%.

The model was also tested with two additional types of inputs with 9 hyperparameter pairs. The test accuracies of all three types of inputs are shown in Figure 2.

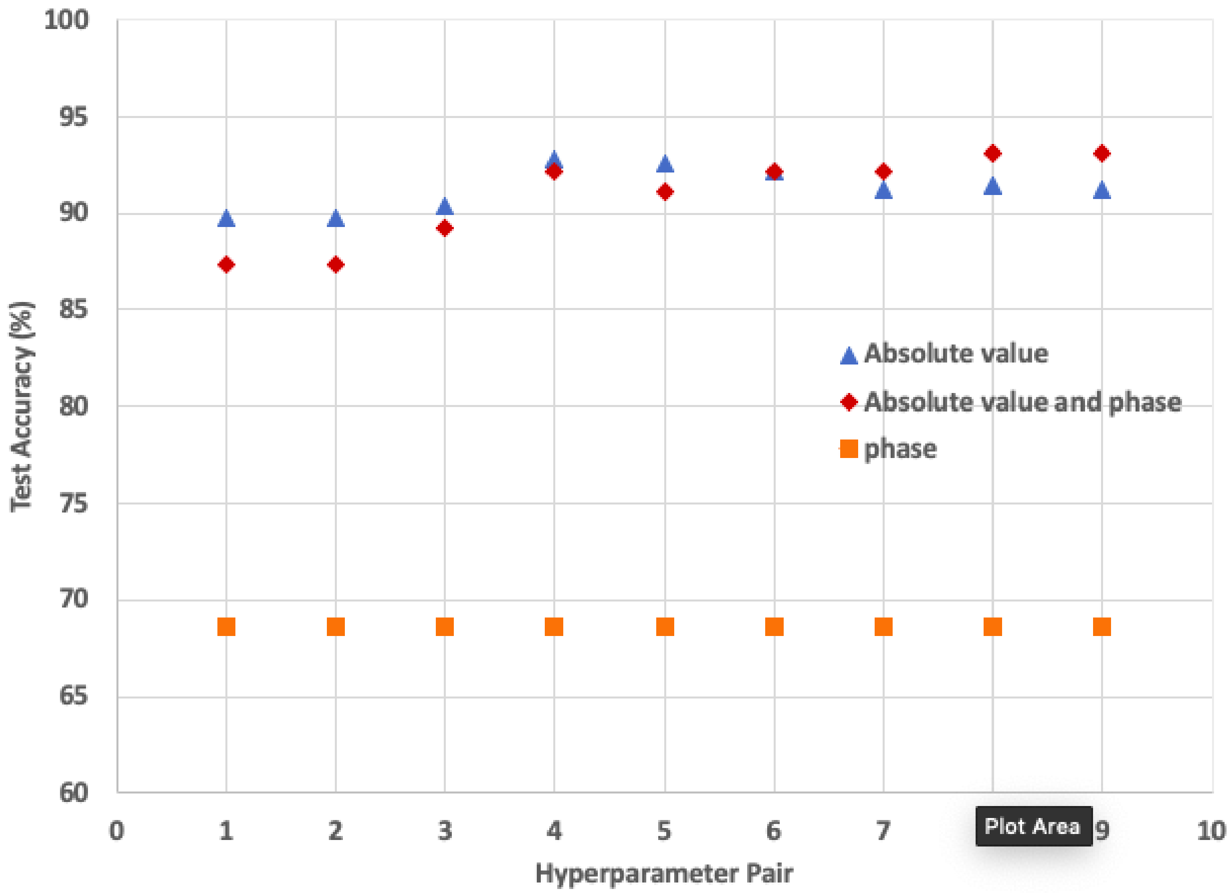

Figure 2 **Testing results of three types of inputs for different hyperparameter pairs. The pairs 1-3 correspond to 128, 256 and 512 nodes in the first fully-connected layer and 0 trainable convolution layer. The pairs 4-6 correspond to 128, 256 and 512 nodes in the first fully-connected layer and 1 trainable convolution layer. The pairs 7-9 correspond to 128, 256 and 512 nodes in the first fully-connected layer and 2 trainable convolution layers.**

The test accuracy of using only the absolute value of spectrograms was around 90% and was used as the reference to compare the results from other inputs. When the model used the combined absolute value and phase information of the spectrograms as inputs, the test accuracy was close to that of using only absolute value. The test accuracy increases as the number of nodes in the first fc layer and the number of trainable convolution layers increase. When the model used only the phase information of the spectrograms as the inputs, the test accuracy was 68.6% and was far below the test accuracy of using the absolute value of spectrograms. The change of test accuracy across different hyperparameter pairs was small.

3.5 Discussion of the results

In this study, a 2D deep learning model was developed from VGG-19 and the spectrograms of the ultrasound RF signal were used as the inputs for the model to diagnose NAFLD. Three different types of images were used as inputs for this model to understand how the model works. The first type of images consisted of only the absolute value of spectrograms. The high accuracy given by the model demonstrated the model's ability to extract useful features from the absolute value of spectrograms of the ultrasound RF signal and related features to the characteristics of NAFLD. The result also suggested that the frequency-dependent information of RF signal contained a lot of information and can be utilized to diagnose diseases and characterize tissue properties. Comparing to use QUS parameter acquired from the phantom-based method, the deep learning approach does not need a physical phantom as the reference to characterize the tissue and can use the frequency information from RF signal for ultrasound diagnosis.

The model was also trained and tested on two additional types of inputs. When inputs consisted of the log-transformed absolute value of spectrogram, the absolute value of spectrogram and the phase information of spectrogram, the model gave similar predictions as using only the absolute value of spectrograms. This similarity can be explained by that only the absolute value of spectrogram contained useful information which helps the diagnosis of NAFLD. The model built from VGG-19 processed the information from each channel independently and only used features extracted from three channels by convolution in the final fully-connected layer. Therefore, the model could use the absolute value of spectrograms from R and G channels and did not use the phase information from the B channel.

When the model was fed only the phase information of spectrograms, the model did not give an accurate prediction and the test accuracy was only 68.6%. The model was more likely to randomly guess the output when given the phase information of spectrograms of RF signal. The low accuracy may suggest that the phase information of spectrograms does not contain useful information for the diagnosis of NAFLD.

There are some limitations to this study. The data used for training and testing were acquired from one single site and by the same machine. Although the test results of the model on the test set were good, the test accuracy of the model on another dataset was unknown because of no more available dataset. Future studies can be done to address the repeatability and reproductivity of using 2D deep learning on ultrasound spectrograms to diagnose NAFLD. In this study, the model was developed from VGG-19. There are many existing 2D deep learning models such as ResNet, Inception, and NASNet. These models can also be used as the base model for transfer learning and gave different results than VGG-19. Future studies can also be done to investigate the performances of theses 2D model on diagnosing NAFLD.

4. Conclusion

This study showed that it is feasible to develop a 2D deep learning model on ultrasound spectrogram of RF signals to accurately diagnose NAFLD. The model can be developed using transfer learning by using existing 2D deep learning models with some hyperparameter optimization. The spectrograms of ultrasound RF signal contained useful information related to the tissue and using deep learning approach on ultrasound RF signal could provide a phantom-free approach for ultrasound diagnosis.

References

[1] R. Loomba and A. J. Sanyal, "The global NAFLD epidemic," *Nature Reviews Gastroenterology and Hepatology*, vol. 10, pp. 686-690, 2013.

[2] Z. M. Younossi, "Non-alcoholic fatty liver disease – A global public health perspective," *Journal of Hepatology*, vol. 70, no. 3. pp. 531–544, 2019.

[3] L. Castera, "Invasive and non-invasive methods for the assessment of fibrosis and disease progression in chronic liver disease," *Best Pract. Res. Clin. Gastroenterol.*, vol. 25, pp. 291-303, 2011.

[4] T. A. Le *et al.*, "Effect of colesevelam on liver fat quantified by magnetic resonance in nonalcoholic steatohepatitis: A randomized controlled trial," *Hepatology*, vol. 56, pp. 922-932, 2012.

[5] M. Noureddin *et al.*, "Utility of magnetic resonance imaging versus histology for quantifying changes in liver fat in nonalcoholic fatty liver disease trials," *Hepatology*, vol. 58, pp. 1930-1940, 2013.

[6] C. W. Hong *et al.*, "Reader agreement and accuracy of ultrasound features for hepatic steatosis," *Abdom. Radiol.*, vol. 44, pp. 54-64, 2019.

[7] M. P. Andre *et al.*, "Accurate diagnosis of nonalcoholic fatty liver disease in human participants via quantitative ultrasound," in *IEEE International Ultrasonics Symposium,* Chicago, IL, 2014, pp. 2375-2377.

[8] E. Uchino, T. Koga, H. Misawa, and N. Suetake, "Tissue characterization of coronary plaque by kNN classifier with fractal-based features of IVUS RF-signal," in *Journal of Intelligent Manufacturing*, vol. 25, pp. 973-982, 2014.

[9] X. Hao, C. Bruce, C. Pislaru, and J. F. Greenleaf, “Classification of normal and infarcted myocardium based on statistical analysis of high-frequency intracardiac ultrasound rf signal,” in *Medical Imaging 2002: Ultrasonic Imaging and Signal Processing*, vol. 4687, pp. 139-146, 2003.

[10] B. Alacam, B. Yazici, and N. Bilgutay, “Breast tissue characterization based on ultrasound RF echo modeling and tumor morphology,” in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439),* Cancun, 2003, pp. 1180-1183 Vol.2.

[11] M. Moradi, P. Abolmaesumi, D. R. Siemens, E. E. Sauerbrei, A. H. Boag, and P. Mousavi, “Augmenting detection of prostate cancer in transrectal ultrasound images using SVM and RF time series,” *IEEE Trans. Biomed. Eng.*, vol. 56, pp. 2214-2224, 2009.

[12] M. L. Oelze and J. Mamou, “Review of quantitative ultrasound: Envelope statistics and backscatter coefficient imaging and contributions to diagnostic ultrasound,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol.63, pp. 336-351, 2016.

[13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” *arXiv e-prints,* 2014.

[14] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning with applications to medical imaging,” *arXiv e-prints,* 2019.

[15] M. Byra *et al.*, “Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images,” *Int. J. Comput. Assist. Radiol. Surg.*, vol.13, pp. 1895-1903, 2018.

[16] D. S. Reddy, R. Bharath and P. Rajalakshmi, "A novel computer-aided diagnosis framework using deep learning for classification of fatty liver disease in ultrasound imaging," *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Ostrava, 2018, pp. 1-5.

[17] N. Tajbakhsh *et al.*, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE Trans. Med. Imaging*, vol. 35, pp. 1299-1312, 2016.

[18] H. G. Kim, Y. Choi, and Y. M. Ro, “Modality-bridge transfer learning for medical image classification,” in *Proceedings - 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2017*, 2018.

[19] C. E. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, “Activation functions: Comparison of Tredns in practice and research for deep learning,” *arXiv e-prints,* 2018.